Using Topical Networks to Detect Editor Communities in Wikipedias

Michael Kretschmer
Advanced Community Information Systems (ACIS), Chair of Computer Science 5 (Information Systems & Databases), RWTH Aachen University
Ahornstr. 55, 52074 Aachen, Germany
kretschmer@dbis.rwth-aachen.de

Bernhard Göschlberger
Research Studio Data Science Research Studios Austria FG
Leopoldskronstr. 30, 5020 Salzburg, Austria
goeschlberger@researchstudio.at

Ralf Klamma
Advanced Community Information Systems (ACIS), Chair of Computer Science 5 (Information Systems & Databases), RWTH Aachen University
Ahornstr. 55, 52074 Aachen, Germany
klamma@dbis.rwth-aachen.de

***Abstract*—The collaboration of Wikipedia editors is well researched, covered by scientific works of many different fields. There is a growing interest to implement recommender systems that guide inexperienced editors to projects which fit their interests in certain topical domains. Although there have been numerous studies focusing on editing behavior in Wikipedia the role of topical domains in this regard is still unclear. In particular, topical aspects of co-authorship are generally neglected. In this paper, we want to determine by which criteria editors usually choose articles they want to contribute to. We analyzed three different language editions of Wikipedia (Vietnamese, Hebrew, and Serbo-Croatian) by building social networks and running community detection algorithms on them, i.e. editors are grouped based on their shared involvement in Wikipedia articles using social network analysis techniques. Then, we related this to the topical domains of these articles based on Wikipedia's user defined category network. Our results demonstrated that communities in Wikipedia tend to edit articles with a higher than average topical relatedness. But the significance and quality of these results vary considerably in the different language versions of Wikipedia. Topical relations between contributors and articles are a complex matter and influenced by a number of different factors, e.g. by culture.**

***Index Terms*—community detection, Wikipedia, social network analysis, topical analysis**

I. Introduction

Wikipedia constitutes one of the largest web collections of encyclopedic information and is among the 10 websites receiving the most traffic. Since a majority of its content is edited and maintained by voluntary users in their free time, Wikipedia can also be interpreted as a social network with users interacting with its content and with each other [1]. However, as an open encyclopedia the quality of its content varies heavily, and certain domains are less developed than others [2]. Additionally, the amount of unfinished articles in a Wikipedia version can be quite significant [3]. These aspects differ between Wikipedia versions of different languages [4]. In 2007 Cosley et al. [5] published their design of "SuggestBot"[1] with the goal of pointing contributors to articles in need of improvement. SuggestBot has since become a helpful tool in supporting editors to find articles to improve [6]. In recent years different designs of recommender systems have been proposed [7], [8] to personalize these suggestions. Morgan and Halfaker identified the sense of community a new Wikipedia editors experiences as an important factor related to the retention rate in a recent report [9]. These subcommunities within Wikipedia are the driving force behind article creation and elaboration [10]. We are therefore interested in analyzing these editor communities and investigate how topical domains relate to communities of Wikipedia contributors across different languages.

Although the research regarding individual author behavior is plenty [11]–[14], community structures among Wikipedia authors and their relation to domains of Wikipedia is mostly unexplored. We propose a general approach based on structural features which is applicable to Wikipedia versions of any language and use it to obtain data from three different Wikipedias. In our approach we combine state-of-the-art methods of constructing a Wikipedia edit network, community detection, and extraction of semantic relationships from the Wikipedia taxonomy. With this we aim to obtain elementary metrics such as size and strength of communities in Wikipedia, as well as their relation to topical domains and possible intersections between them. Results obtained by Halatchliyski et al. [12] indicate that knowledge domains in Wikipedia are not exclusive. Thus, we also compare results obtained from overlapping and non-overlapping communities.

In the following section we are exploring some of the previous research that relates to our investigation. Afterwards, we present our method of grouping contributors to Wikipedia and computing a meaningful measure of topical relatedness between them. We then apply this method on three versions of Wikipedia in different languages and discuss our findings. Lastly, we offer a conclusion of the obtained results and some ideas for future research.

II. Related Work

Wikipedia has been the subject of many scientific works of different domains and with different goals in mind. In

[1] https://en.wikipedia.org/wiki/User:SuggestBot

this section we present previous scientific work that is most relevant to our research. This can loosely be categorized under three aspects.

A. Behavioral Analysis

The fact that Wikipedia editors do not choose articles to contribute to randomly has been shown by studies like Keegan et al. [10]. They employed statistical analysis and found that certain articles were edited by groups of authors with different experience levels. They found that the experience of authors influences the domain and volume of their contributions.

In 2012 Wikipedia launched its Teahouse project[2] with the intent of studying the effectiveness of an approach of social inclusion on newcommers' retention rates. The idea behind the Wikipedia Teahouse was to enable newcommers to make first meaningful contributions and integrate them into the Wikipedia community [6]. This sense of community relates to findings of Welser et al. [15] that groups of Wikipedia contributors form a social network. This measure significantly improved the retention rate of unexperienced editors as found by Morgan and Halfaker [9]. On the other hand they also stated that the system by which contributors are referred to projects could be improved, especially considering those with no or very little recorded activity. A central issue in this regard is assessing editors' interests in order to make effective recommendations. These recommendations are not limited to specific articles, but domain centric intended to guide newcommers to editor communities where they are introduced to the basics of Wikipedia editing [6]. Personalizing recommendations based on interest generally increases the likeliness that an editor accepts the recommendation [5], [7]. It has however also been show that interest is a difficult aspect to model in Wikipedia [11], [16].

In an early approach Turek et al. [11] investigated teams of Wikipedia contributors who collaborated on an article. They identified a number of metrics correlated to the quality of these articles. Most relevant to our work are the results Turek et al. observed regarding the topical aspect of editor teams. Here, they found that teams of contributors with less collective experience in a certain domain on average produce articles of higher quality. The proposed explanation referred to the granularity of Wikipedia categories and that more mature articles are typically tagged with a higher amount of categories. To avoid this circumstance, we rely on a path based metric over Wikipedia's category network proposed by Chernov et al. [17] which has been found to reasonably approximate human judgment [18].

B. Wikipedia Categories

Wikipedia categories are organized in a hierarchical network with broad high level categories splitting up into increasingly specific categories. The hierarchy is implemented based on category tags assigned to category pages the same way categories are assigned to articles. Thus, the hierarchy is not enforced and cycles are possible in the network, especially considering that most categories are assigned manually. Munchnik et al. [19] performed a comprehensive analysis of the categories of a number of different Wikipedias. They found the number of cycles in the Wikipedia versions they investigated to be remarkably low. Thus, the overall structure of Wikipedia's category network exhibits clear hierarchical properties. Furthermore, Muchnik et al. showed that article relations extracted from e.g., intra-wikilinks[3] are structurally similar to the category hierarchy. As a consequence, the category network also holds a great potential for retrieval of semantic information.

Strube and Ponzetto [20], [21] showed that the semantic relatedness of terms that can be inferred based on the Wikipedia category network is comparable in quality to other established systems such as Google and WordNet. Schönhofen [22] used this category information among other properties of Wikipedia articles to build a system for automatically detecting the topic of an arbitrary document. Even without any content analysis of the article body, the proposed method achieved up to 86% accuracy.

C. Communities in Wikipedia

Community detection is frequently used in social network analysis to find users with similar behavior [23]. The methods by which actors in a network are grouped influence what kinds of communities are detected [24]. Thus, the chosen method is dependent on the goal with which community detection is employed. A common approach of grouping Wikipedia authors utilized in many other works [4], [11]–[14] is to collect all contributors of certain sets of articles. In these works links between authors and articles they edited are usually weighted based on the size and amount of contributions.

What many of these works lack is a basic description of editor communities found in Wikipedia. General metrics regarding such communities (e.g., size, activity, etc.) are not in the scope of these and similar works. Our goal is it to obtain basic quantitative information regarding the differences of author communities across multiple versions of Wikipedia. For this, we rely on the Speaker-Listener Label Propagation algorithm (SLPA) [25] which is sensitive to the underlying network structure [26].

III. Methodology

The implementation of a topical relatedness metric is a multi-layered process. We constructed networks of authors and articles based on Wikipedia's edit history on which we performed community detection. From the resulting author communities we collected pairs of edited articles and computed their distance in the category taxonomy. The resulting sets of path lengths were then statistically evaluated. The individual tasks including important design decisions are described in this section. The source code of our parsing program along

[2]https://meta.wikimedia.org/wiki/Research:Teahouse

[3]https://en.wikipedia.org/wiki/Hyperlink#Wikis

with the obtained network data is available on GitHub[4] and Sciebo[5].

A. Building the Networks

For the purpose of our analysis we required a number of different networks. The data needed to construct these networks is freely available for anyone to download[6]. Our approach is based on three different types of dump files: an index file listing every Wikipedia article and category page, a collection of history files detailing the contributions made to every article, and a categorylinks file linking Wikipedia pages to categories.

We preprocessed these files using techniques similar to other works [11], [12], [14], [27]. We excluded contributions of non-registered users which were identified by them being associated with an IP address rather than a username. Edits which were made by users who are verified bot accounts were excluded as well. We also did not consider reverts as contributions which we identified by comparing hashes over the article body. Furthermore, we wanted to exclude contributions, characterized by a lack of topical information regarding their content. We achieved this by disregarding any contribution tagged with the *minor* flag as described in [12]. Wikipedia states that these kinds of edits "differ only superficially (typographical corrections, etc.), in a way that no editor would be expected to regard as disputable"[7]. We additionally made an effort to remove redirect and disambiguation pages from the set of articles. This was done by removing all pages tagged with a redirect or disambiguation category. Thus, we obtain our sets of articles V_{Art}, authors V_{Auth} and revisions $E_{Rev} = \{(art, auth) \mid art \in V_{Art},\ auth \in V_{Auth}\ and\ auth\ made\ an\ contribution\ to\ art\}$.

The *history network* is formally defined as a bipartite graph (V_H, E_{Rev}) where $V_H = V_{Art} \cup V_{Auth}$. In this network an edge between an author and an article is equivalent to a record in the history dump file, thus there may be multiple edges between the same two nodes. Edge weights as defined in other works [11], [14], [28] were not computed. We constructed two slightly different versions of this network. The *directed history network* featuring directed edges pointing from an article to an author and the *undirected history network* where edges are undirected.

The *author network* features similar information as the history networks but is much smaller. Formally it is defined as a undirected graph (V_{Auth}, E_H) where $E_H = \{(v, w) \mid v, w \in V_{Auth}\ and\ \exists(a, v),\ (a, w) \in E_{Rev}\ with\ a \in V_{Art}\}$. Thus, edges connect authors that edited at least one mutual article.

The *category network* is a directed graph made up of Wikipedia article and category pages and links between them denote category assignments. Edges in this network are directed from the tagged page to the assigned category. We performed a topical filtering on this network with the goal of eliminating non-topical paths between article pairs. For this, a number of large container categories which are based on structural rather than semantic criteria was manually selected. Ponzetto and Strube [21] used a similar method in order to construct their semantic taxonomy. Their criterion for a non-topical category was based on certain strings being part of the category title. In our approach, all pages tagged with one or more of the following categories were considered non-topical: **Hidden categories**, **Wikipedia template categories**, **Wikipedia maintenance**, **Tracking categories**, **Stub categories**, and **All redirect categories**.

B. Community Detection

The algorithm we used to detect author communities is the Speaker-Listener Label Propagation Algorithm (SLPA) [25]. SLPA is an overlapping community detection (OCD) algorithm that works well even on large scale graphs and also produces reliable results on bipartite networks [26]. Thus, it constitutes a good fit for our approach. A maximum amount of iterations can be specified so the algorithm terminates before the condition for convergence is satisfied. We were not able to find a general recommendation how many iterations SLPA should perform other than "more than 20" [26], so we ran the algorithm 3 times. Once with 3, 10, and 100 iterations respectively. As an additional test the algorithm was run with 1,000 max iterations but our trials showed that at least on a superficial level performing 100 or 1,000 iterations does not produce substantially different results. It is also possible to limit the amount of communities a vertex is allowed to be part of. It should be noted that doing so only influences the final result. In intermediate steps of the algorithm, every vertex is labeled with multiple communities. We applied SLPA three times capping the maximum amount of communities to one so that only the community with the strongest affiliation was displayed. This way we obtained three collections of non-overlapping communities for the two history and the author networks respectively. We also performed 10 and 100 iterations of SLPA on the undirected history network restricting the community count of each node to 20. Thus, we additionally obtained two overlapping community collections.

In order to assess the general quality of communities detected by our approach we computed Newman's modularity [29] over the communities within the undirected history network and the author network. Modularity may be used as a general indication of the expressiveness of communities. The original formula of modularity given in (1) assumes an undirected network and exclusive communities.

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \tag{1}$$

In the formula m denotes the degree of the network, k_i the degree of vertex i, and A is the adjacency matrix. $\delta(c_i, c_j)$ equals 1 if vertices i and j are part of the same community and

[4]https://github.com/rwth-acis/Topical-Analysis-Wikipedias
[5]https://rwth-aachen.sciebo.de/s/VgtWe8NJ7BV3ON5
[6]https://dumps.wikimedia.org/
[7]https://en.wikipedia.org/wiki/Help:Minor_edit

```
A,B = getRandomArticleFromCommunity
catsA = categories(A)
catsB = categories(B)
for each c_a in catsA do
  for each c_b in catsB with c_b != c_a do
    path = shortestPath(c_a, c_b)
      if (path not in database)
         storeToDatabase(path)
  done
done
```

Fig. 1. Pseudocode of our algorithm for computing topcial relatedness

0 otherwise. We used formula 2 taken from [30] to compute modularity of the directed history network.

$$Q = \frac{1}{2m}\sum_{ij}(\frac{A_{ij}}{m} - \frac{k_i^{out}k_j^{in}}{m^2})\delta(c_i, c_j) \qquad (2)$$

k_i^{out} and k_i^{in} in this case denote the in-degree and out-degree of node i. Since the computation is rather expensive for large networks we did not implement an additional modularity of overlapping communities.

C. Topical Relatedness

The main part of our analysis was the computation of a topical relatedness metric for the previously obtained communities. For each community we collected the sets of articles, members of the respective community contributed to. Our topical relatedness metric was then computed over these sets and compared to the values obtained for random sets of articles. Specifically, we sampled random pairs of articles from within the community and saved the categories of both articles to separate sets. We then computed the shortest path in the category network for every pair of categories from the separate sets and stored the results in our database. Kittur et al. [18] found simple edge counting as a distance measure to be rather robust and substantially similar to more elaborate metric definitions *e.g.,* "normalizing by taxonomy depth". Thus, we also employed this simple definition of distance regarding our measure of topical relatedness. Pseudocode for this algorithm is given in Figure 1. For our null model we sampled arbitrary pairs of articles. An important distinction in this regard has to be made concerning the set of articles Art_{Commy} from which articles A and B are chosen. For the computation of topical relatedness of non-overlapping communities the set is defined as all articles that were edited by members of the community: $Art_{Commy} = \{art \in V_{Art} \mid \exists(art, auth) \in E_{Rev}\, with\, auth \in Commy\}$. In the case of overlapping community detection we sampled from the articles that are part of the community: $Art_{Commy} = \{art \in V_{Art} \mid art \in Commy\}$.

IV. Experimental Results

We performed the aforementioned approach on three different versions of Wikipedia, each in a different language.

vi	History Graph	Author Graph	Page Graph
Authors	46,236	46,236	-
Articles	442,558	-	1,584,062
Categories	-	-	208,752
Edges	2,323,485	3,462,238	4,159,548
he	**History Graph**	**Author Graph**	**Page Graph**
Authors	32,238	32,238	-
Articles	111,946	-	456,988
Categories	-	-	60,435
Edges	1,588,576	2,093,581	1,729,619
sh	**History Graph**	**Author Graph**	**Page Graph**
Authors	7,558	7,558	-
Articles	489,553	-	533,725
Categories	-	-	44,172
Edges	728,914	70,705	1,509,010

Fig. 2. Network Sizes.The History Network in this case only contains articles for which edits are present in the history file

The languages featured are **vietnamese (vi)**, **hebrew (he)**, and **serbo-croatian (sh)** mainly on the basis of their respective *depth* value. Depth of a Wikipedia version indicates how commonly the content of Wikipedia is updated and gives a rough estimate of its quality. Furthermore, these versions are diverse in terms of ratios between authors, active authors, articles, and revisions. We also attempted to apply our method on the far more developed English Wikipedia, but this failed due to limited computational resources.

A. Networks

The constructed networks exhibited noticeable differences across the investigated versions of Wikipedia. This can in large parts be explained by the different ratios of articles compared to authors. We can also compare the obtained values for the networks to Wikipedia's metrics regarding these languages[8]. From this we see that only between 2% (sh) and 7% (he) of all edits and around 6% of all users are part of our network. For the ratio of articles we get different results ranging from about 37% (vi) to 109% (sh). The additional articles in the Serbo-Croatian history network are likely disambiguation or redirect pages which are not considered to be articles[9]. Manually created disambiguation/redirect pages might not be tagged properly which causes them not to be detected by our approach. This, we assume is also the reason that the number of articles in our dataset is between roughly 120% (sh) and 190% (he) of what it should be based on Wikipedia's data. The filtering of non-topical pages proved to be difficult. We were for example unable to find any large Hebrew container category of redirection pages. Our approach filtered about 15,000 articles and categories from the Hebrew Wikipedia and we excluded around 34,000 and 175 pages from the Vietnamese and Serbo-Croatian versions respectively. Because our approach of topical filtering produced such poor results on the Serbo-Croatian data, we instead included only those articles which were edited by authors in our set.

[8]https://en.wikipedia.org/wiki/List_of_Wikipedias#Detailed_list
[9]https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article?

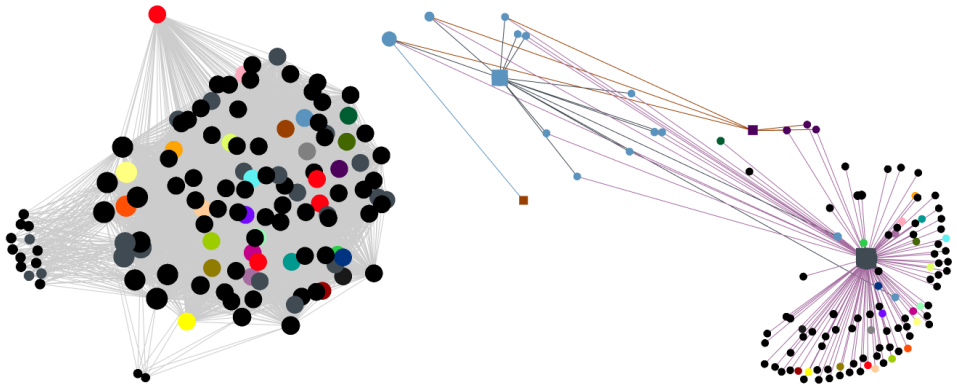

Fig. 3. Excerpt of the Vietnamese author network (left) and undirected history network (right) featuring 4 articles (squares) and 126 authors (circles) with nodes colored by community

Collection	Community Count	Avg Article Count	Avg Author Count	Max Article Count	Max Author Count
Vietnamese					
DH3	46,236	19.602	1	42,588	1
DH10	26,164	33.633	1.767	42,588	78
DH100	26,164	33,633	1.767	42,588	78
UH3	38,884	22.434	1.189	42,588	178
UH10	33,784	24.117	1.369	114,890	759
UH100	31,630	24.853	1.462	98,743	1,052
OH10	47,425	1.773	1.335	7,402	1,186
OH100	39,628	1.866	1.248	7,317	1,255
AN3	7,082	98.6415	6.529	146,053	8,544
AN10	6,346	72.109	7.286	433,196	31,728
AN100	5,827	76.843	7.935	439,624	39,905
Hebrew					
DH3	29,277	17.722	1.101	14,574	79
DH10	24,698	20.613	1.305	14,574	141
DH100	24,698	20.613	1.305	14,574	141
UH3	31,898	16.363	1.011	14,608	24
UH10	22,254	18.299	1.449	54,042	655
UH100	20,840	18.618	1.547	55,399	799
OH10	35,186	1.639	1.326	4,117	560
OH100	26,315	1.799	1.327	4,119	599
AN3	19,941	20.228	1.617	63,473	3,079
AN10	1,687	69.075	19.110	111,572	27,078
AN100	1,487	76.210	21.680	111,660	30,639
Serbo-Croatian					
DH3	7,511	75.344	1.006	228,796	4
DH10	7,085	79.802	1.067	228,796	14
DH100	7,085	79.802	1.067	228,796	14
UH3	7,474	75.580	1.011	228,796	65
UH10	5,761	92.914	1.312	229,424	1,133
UH100	5,398	98.976	1.400	229,714	1,382
OH10	9,630	42.746	1.274	227,252	1,400
OH100	5,625	71.640	1.409	227,199	1,405
AN3	6,475	85.937	1.167	306,779	997
AN10	1,165	420.512	6.487	488,171	6,273
AN100	1,138	430.345	6.641	488,225	6,385

Fig. 4. Metrics of the community collections detected in the directed and undirected history network (DH and UH), in the author network (AN), and using OCD (OH) by either 3, 10, or 100 iterations of SLPA

The ratio of articles to edges gives an indication how often an article is revised on average. This gives us about 1.5, 5, and 14 revisions for the Serbo-Croatian, Vietnamese, and Hebrew version respectively.

B. *Communities*

Figure 3 gives an idea of the nature of the created networks and the detected communities. We can see that the author network is much denser and even nodes from differing communities are clustered together more closely. As listed in Figure 4, we obtained the same metrics for 10 and 100 iterations of SLPA on the directed history network every time. This circumstance is most likely due to the network being directed with edges always pointing from pages to author nodes. Therefore, only author nodes change their community and only based on random factors like the sequence in which communities are updated or how ties between equally frequent communities are broken. Additionally, a high ratio of communities featuring only one author or only one mutually edited article persists across all investigated languages. This value is strictly over 52% for the history networks and even over 64% for the author network across all languages. Communities in the author network are heavily unbalanced with a single community containing as much as 99.7% (he *AN100*) of all authors. Looking at the communities detected in the directed history network the Serbo-Croatian version stands out due to the small number of extremely active authors. Further investigation revealed two extraordinarily active contributors, with 233,106 and 169,360 edits on record. Based on the discussions on one of these authors user page[10] it seems that some of these edits may have been automatic.

The quality of the detected communities is generally similar for all investigated versions with regard to their modularity as shown in Figure 5. The communities produced by performing SLPA on the directed history network are very close to equivalent to random communities. An outlier is given by the Serbo-Croatian undirected history network. The modularity of communities detected within this network is much higher than for any other communities we detected in the course of our work. What we can clearly see is that more iterations of SLPA result in more modular communities, which is why we computed the topical relatedness for communities detected with 100 iterations.

[10]https://sh.wikipedia.org/wiki/Razgovor_sa_korisnikom:Dcirovic

lng	DH3	10	UH3	10	100	AN3	10	100
vi	0.002	0.009	0.072	0.127	0.145	0.002	0.269	0.307
he	0.002	0.012	0.068	0.138	0.159	0.040	0.273	0.301
sh	0.001	0.004	0.523	0.750	0.753	0.013	0.233	0.238

Fig. 5. Modularity of detected communities

Languages	Arbitrary	DH100	UH100	OH100	AN100
Vietnamese	8.518	6.392	6.633	5.856	7.712
Hebrew	7.781	7.258	7.395	6.756	7.806
Hebrew No TF	7.335	6.994	7.071	-	7.312
Serbo-Croatian	7.660	6.702	7.038	6.111	7.626

Fig. 6. Means of category path lengths between articles sampled from community collections compared to arbitrary article pairs. Hebrew No TF stands for no topical filtering of pages. Green values are statistically significant at a 99% significance level, red values are not at a 95% significance level

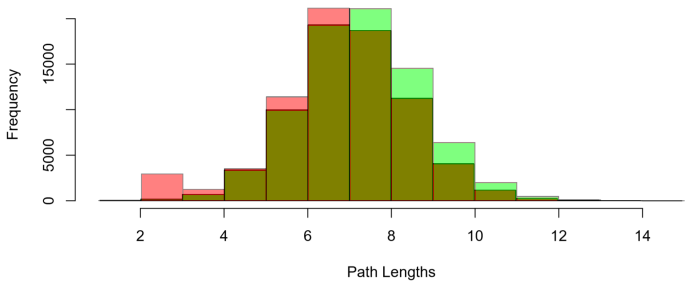

Fig. 8. Superimposed histograms of path lengths computed in topically filtered (green) and unfiltered (red) Hebrew Wikipedia. Frequency of the topically unfiltered path lengths is normalized to fit into the same chart

C. Topical Relatedness

Figure 6 shows that the path lengths connecting pairs of articles sampled from detected communities are on average shorter than to those sampled from the entire set of articles. The Wilcoxon signed rank test [31] shows that almost every result is statistically significant at a 99% confidence level. Results obtained from the communities detected in the history networks strictly indicate a stronger topical relatedness of articles than those detected in the corresponding author network.

The mean differences as computed by the Wilcoxon signed rank test at a 99% confidence level are close to the mean differences listed in Figure 6 for the Vietnamese and Serbo-Croatian version. For the Hebrew communities the mean difference is at most 0.1 for communities detected in the directed history network and in the range of 10^{-5} for the other communities. We can also see that articles worked on by the communities detected in the author network are on average less topically related, although as stated the mean difference is very low. If no topical filtering of Wikipedia pages is performed the mean differences are even lower with a maximum of $7.395 * 10^{-6}$. Figure 8 depicts the results of the shortest path computations for categories of arbitrary Hebrew articles. We see that the average path length computed in the topically unfiltered Hebrew category network is shorter due to the higher ratio of paths shorter than seven. The Wilcoxon signed rank test verified that these differences are statistically significant at a 99% confidence level. Performing a goodness of fit test on the computed path lengths, we found that they follow neither a normal nor a Poisson distribution.

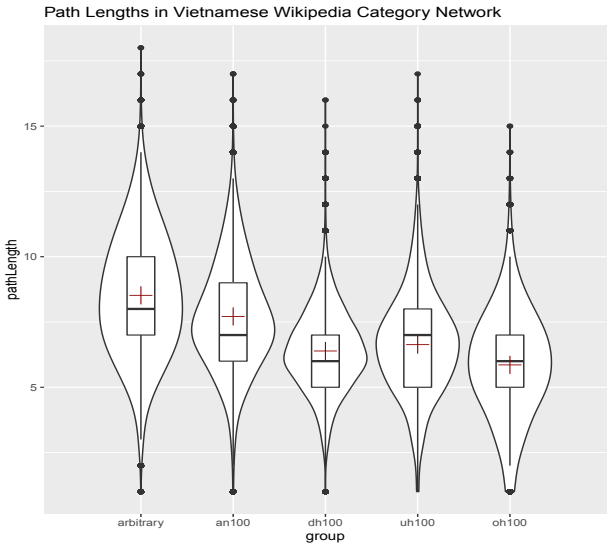

Fig. 7. Violin Plots of Vietnamese path lengths: arbitrary, author network, directed history network, undirected history network, overlapping communities (left to right). The plots include Boxplots and distributions with the red cross indicating the average and the dots signaling outliers

D. Overlapping Communities

We can clearly see from Figure 4 that the average number of articles within a community is much lower than the average number of articles which were edited by communities (compare **Avg Article Count** between collections **UH10/100** and **OH10/100**). The average article count based on community affiliation is in fact comparable to the average author count. The exception being the Serbo-Croatian Wikipedia version. Path lengths between articles sampled from overlapping communities are on average much shorter than for arbitrary article pairs. The mean difference is also significantly higher than that computed for non-overlapping communities at a 99% significance level. We also looked at the three most common categories for both the largest overlapping author and article communities. The results are displayed in Figure 9. Noticeable is that the communities with the most articles only feature 79 and 156 distinct authors for the Serbo-Croatian and Hebrew version respectively. The largest author communities on the other hand feature around 600 and 1,400 authors respectively.

V. Discussion

Our results show that author communities in Wikipedia tend to edit articles with a higher than average topical relatedness. We also demonstrated that the strength of topical relatedness to a high degree dependents on the construction of the network in which we perform the community detection. The topical relatedness of communities detected within our author network was significantly lower for all investigated versions of Wikipedia. This is a logical consequence of the fact that such author communities on average feature considerably more

Community	First	Second	Third
Vi articles/authors	Windows games	Living people	Manga Series
He articles	Personalities of the fifth aliyah	People buried in the Kiryat Shaul cemetery	Jews buried in Har Hamenuhot
He authors	English-language film	American film and TV actors	American movies
Sh articles	Communes of the Province of Turin	... Cuneo	... Vicenza
Sh authors	Living People	Drama Films	American Films

Fig. 9. Most central categories for the largest author and article community. Centrality being defined based on the number of times a certain tag showed up

articles than those detected in the history networks. The results of the community detection we performed, demonstrate that community structures among Wikipedia contributors can be identified based on their edit history. We computed a simple modularity metric to assess the strength of these communities and found them to be rather weak in most cases. However, previous research indicates that this modularity measure is less sensitive to networks with especially small communities [32], [33]. The author networks are highly connected as can be seen in Figure 3 which also causes the modularity to be lower. In that regard, we observed that the majority of communities only features one author or one article which is related to the Matthew effect or "rich-get-richer" rule. The prevalence of preferential attachment in Wikipedia is a well studied phenomena [34]. This causes power law distributions [1], [35] in the history network and thus leads to some very large and many small communities. Research of Petrushyna et al. [4] found that a majority of articles are created by only a relatively small amount of contributors. These factors make link-based approaches (*e.g.,* collaborative filtering) of assessing author behavior very challenging, since the information gained from it is severly limited for the vast majority of contributors.

Petrushyna et al. analysis also illustrates that different versions of Wikipedia display significant culturally founded types of behavior and interaction. Our results confirm these findings and complement them with a topical dimension. The impact of topical domains on the editing behavior of author communities is not uniform throughout all Wikipedia versions. These cultural distinctions also have to be considered while collecting data. The Serbo-Croatian Wikipedia which features an extremely high number of redirection pages exemplifies this. Previous research of Wikipedia redirects [36]–[38] indicates that this is due to the pluricentric[11] nature of the Serbo-Croatian language.

Results we obtained by performing overlapping community detection suggest that there is a considerable difference between communities of authors and communities of article nodes. List type categories that exist in a specific topical dimension and directly link to a high number of articles have good qualifications to be at the center of very large article communities. Large author communities on the other hand tend to form based on interest and around more generalized categories that split up into more specific sub communities. The Hebrew Wikipedia with its article community centered around specific groups of people and its author community strongly connected to American film and television exemplifies this very well. It has been shown that especially categories related to individuals constitute a significant outlier regarding their frequency [2], [18], [35]. In terms of social network analysis, this is an important insight, since it shows that author and article nodes are interdependent. Our computed path lengths also illustrate that these topical domains are overlapping and not exclusive.

Jankowski et al. [16] were unable to find a strong basis for social network interpretations of Wikipedia contributions as defined by Turek et al [11]. They collected metrics similarly to Turek et al. and compared the results with data from a survey they had Wikipedia users fill out. This shows that the interpretation of behavioral data of Wikipedia users is a complex issue and should always be viewed critically. Aspects like non-topical categories and contributions not founded by specific topical interest are a major influence regarding such data. We showed that performing a filtering of non-topical categories has a statistically significant influence on the topical relatedness we computed for the Hebrew Wikipedia. On the other hand, our topical filtering did not perform well on the Serbo-Croatian version. These results illustrate the limitations of our approach.

1) Our value of topical relatedness is given by the mean difference of path lengths connecting articles which are associated to author communities. This is a very abstract value which is hard to interpret in a semantic manner.
2) Our approach of topical filtering disregards the majority of authors and contributions from the dataset.
3) The results of our topical filtering are volatile and the non-topical categories have to be selected manually.

The general issue is that Wikipedia data is noisy [7], [8], which makes exact evaluations on such a diverse set of data difficult.

VI. Conclusion and Future Work

In conclusion, we found a stronger topical relation between author communities within Wikipedia than expected by pure chance. We also demonstrated that these communities are non-exclusive, and that article and author communities display significant structural differences. We achieved this result by exploiting the network characteristics of Wikipedia history and category information using network analysis procedures. This way, we were able to group Wikipedia authors and describe community structures within Wikipedia and their topical relation in a way that has not been implemented before. Wikipedia however constitutes a complex system of networks which need to be modeled and analyzed with particular care. This is due to its open and decentralized nature which can cause inconsistencies and in parts complicate the analysis

[11] https://en.wikipedia.org/wiki/Pluricentric_language

procedures. Additionally, the structures of these networks display significant differences across Wikipedia versions of different languages. Therefore, it is also important to consider the unique characteristics of the investigated versions.

We identified the power law distribution of links in the history network as a major challenge regarding the analysis of author contributions. Another challenging task in this regard is identifying which links between author and articles as well as articles and categories are topically meaningful. In this regard, results obtained by our method of topical filtering demonstrate that it does not yet provide a satisfactory level of robustness and needs to be improved upon.

Still, we were able to show that author communities are correlated to Wikipedia's categories. Based on this result, we can ask further questions regarding the topical relations of author communities. Questions like which factors influence the strength of topical relatedness, which intersections exist between certain domains, and how these aspects have changed over time.

REFERENCES

[1] R. Klamma and C. Haasler, "Dynamic network analysis of wikis," in *Proceedings of I-Know*, vol. 8, 2008, pp. 21–22.

[2] A. Halavais and D. Lackaff, "An analysis of topical coverage of wikipedia," *Journal of Computer-Mediated Communication*, vol. 13, no. 2, pp. 429–440, 2008.

[3] S. Banerjee and P. Mitra, "Filling the gaps: Improving wikipedia stubs," in *Proceedings of the 2015 ACM Symposium on Document Engineering*. ACM, 2015, pp. 117–120.

[4] Z. Petrushyna, R. Klamma, and M. Jarke, "The impact of culture on smart community technology: The case of 13 wikipedia instances." *IxD&A*, vol. 22, pp. 34–47, 2014.

[5] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl, "Suggestbot: Using intelligent task routing to help people find work in wikipedia," in *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 2007, pp. 32–41.

[6] J. T. Morgan, S. Bouterse, H. Walls, and S. Stierch, "Tea and sympathy: Crafting positive new user experiences on wikipedia," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 839–848.

[7] E. Wulczyn, R. West, L. Zia, and J. Leskovec, "Growing wikipedia across languages via recommendation," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 975–985.

[8] R. Yazdanian, L. Zia, and R. West, "The elicitation of new users interests on wikipedia," in *Wiki Workshop*, 2018.

[9] J. T. Morgan and A. Halfaker, "Evaluating the impact of the wikipedia teahouse on newcomer retention," 2018.

[10] B. Keegan, D. Gergle, and N. Contractor, "Do editors or articles drive collaboration?: Multilevel statistical network analysis of wikipedia coauthorship," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 427–436.

[11] P. Turek, A. Wierzbicki, R. Nielek, A. Hupa, and A. Datta, "Learning about the quality of teamwork from wikiteams," in *IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust*. IEEE, 2010, pp. 17–24.

[12] I. Halatchliyski, J. Moskaliuk, J. Kimmerle, and U. Cress, "Explaining authors' contribution to pivotal artifacts during mass collaboration in the wikipedia's knowledge base," *International Journal of Computer-Supported Collaborative Learning*, vol. 9, no. 1, pp. 97–115, 2014.

[13] T. Iba, K. Nemoto, B. Peters, and P. A. Gloor, "Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6441–6456, 2010.

[14] J. Lerner and A. Lomi, "Diverse teams tend to do good work in wikipedia (but jacks of all trades don't)," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 214–221.

[15] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith, "Finding social roles in wikipedia," in *Proceedings of the 2011 iConference*. ACM, 2011, pp. 122–129.

[16] M. Jankowski-Lorek, S. Jaroszewicz, Ł. Ostrowski, and A. Wierzbicki, "Verifying social network models of wikipedia knowledge community," *Information Sciences*, vol. 339, pp. 158–174, 2016.

[17] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantics relationships between wikipedia categories," *SemWiki*, vol. 206, 2006.

[18] A. Kittur, E. H. Chi, and B. Suh, "What's in wikipedia?: Mapping topics and conflict using socially annotated category structure," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 1509–1512.

[19] L. Muchnik, R. Itzhack, S. Solomon, and Y. Louzoun, "Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies," *Physical Review E*, vol. 76, no. 1, 2007.

[20] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *AAAI*, vol. 6, 2006, pp. 1419–1424.

[21] S. P. Ponzetto and M. Strube, "Deriving a large scale taxonomy from wikipedia," in *AAAI*, vol. 7, 2007, pp. 1440–1445.

[22] P. Schönhofen, "Identifying document topics using the wikipedia category network," *Web Intelligence and Agent Systems: An International Journal*, vol. 7, no. 2, pp. 195–207, 2009.

[23] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 115–135, 2016.

[24] Y. Zhang, J. Wang, Y. Wang, and L. Zhou, "Parallel community detection on large networks with propinquity dynamics," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 997–1006.

[25] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven label propagation algorithm," in *2011 IEEE Network Science Workshop*. IEEE, 2011, pp. 188–195.

[26] J. Xie, B. K. Szymanski, and X. Liu, "Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 344–349.

[27] X. Li, J. Tang, T. Wang, Z. Luo, and M. De Rijke, "Automatically assessing wikipedia article quality by exploiting article–editor networks," in *European Conference on Information Retrieval*. Springer, 2015, pp. 574–580.

[28] D. Jurgens and T.-C. Lu, "Temporal motifs reveal the dynamics of editor interactions in wikipedia," in *The Sixth International AAAI Conference on Weblogs and Social Media*, vol. 6. AAAI, 2012.

[29] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, 2004.

[30] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity," *New Journal of Physics*, vol. 9, no. 6, p. 176, 2007.

[31] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[32] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[33] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1275–1276.

[34] A. Capocci, V. D. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia," *Physical review E*, vol. 74, no. 3, 2006.

[35] J. Voss, "Measuring wikipedia," in *International Conference of the International Society for Scientometrics and Informetrics*, vol. 10, 2005.

[36] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and accessing wikipedia as a lexical semantic resource," *Data Structures for Linguistic Resources and Applications*, vol. 197205, 2007.

[37] O. Medelyan, I. H. Witten, and D. Milne, "Topic indexing with wikipedia," in *Proceedings of the AAAI WikiAI workshop*, vol. 1, 2008, pp. 19–24.

[38] D. N. Milne, I. H. Witten, and D. M. Nichols, "A knowledge-based search engine powered by wikipedia," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 445–454.